The ethics of fighting disinformation

Summary

Authors: Marko Kovic / Adrian Rauchfleisch

The global rise of digital disinformation has prompted academia, policymakers and civil society to develop and deploy interventions against disinformation. However, such interventions can also cause damage by restricting the very principles of deliberative democracy they seek to protect. This benefit-vs.-harm conundrum poses an important ethical challenge: How much intervention harm is too much? In this paper, we develop an analytical framework for evaluating the ethical status of disinformation interventions. We proceed in four steps. First, we propose a taxonomy of disinformation interventions. Second, we discuss the available evidence for the effectiveness of the various intervention types. Third, we evaluate the potential damage of disinformation interventions from a consequentialist perspective. Fourth, we combine our findings in a framework that weighs effectiveness against risk. We argue that the group of high net-benefit interventions is ethically unobjectionable, whereas the group of high-impact high-damage interventions, which can be thought of as deliberative weapons of mass destruction, should be used with great restraint. The main benefit of our proposed framework is that it is not a static one-time assessment but rather a generalized and dynamic tool that can be updated with future research: As the evidence on intervention impact grows and becomes more precise, so do the ethical assessments generated with the framework.